Environments for reproducibility



Javier Moldón

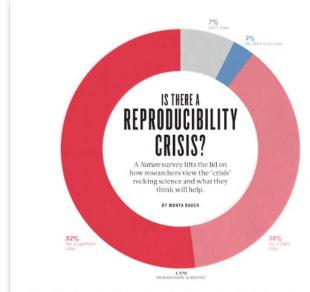
IAA-CSIC (Granada)

Herramientas para la reproducibilidad del análisis científico Curso del CSIC April 20, 2022





"Reproducibility crisis" in the sciences? (Baker 2016, Nature 533, 452)



ore than 70% of researchers have tried and failed to who took a brief online questionnaire on reproducibility in research. my experience wasn't uncommon."

The data reveal sometimes-contradictory attitudes towards reproducibility. Although 52% of those surveyed agree that there is a significant published work. Being at the cutting edge of science means that some-'crisis' of reproducibility, less than 31% think that failure to reproduce times results will not be robust, says Munafo. "We want to be discovering published results means that the result is probably wrong, and most say new things but not generating too many false leads." that they still trust the published literature.

Data on how much of the scientific literature is reproducible are rare THE SCALE OF REPRODUCIBILITY and generally bleak. The best-known analyses, from psychology' and But sorting discoveries from false leads can be discomfitting. Although chemists generally showing the most confidence.

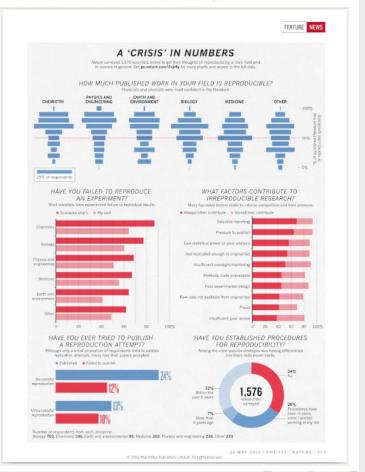
be." But just recognizing that is a step forward, he says. "The next step revealing too much about their own projects. may be identifying what is the problem and to get a consensus."

Failing to reproduce results is a rite of passage, says Marcus Munafo, a reproduce another scientist's experiments, and more biological psychologist at the University of Bristol, UK, who has a longthan half have failed to reproduce their own experiments. Those are some of the telling figures that he says, "I tried to replicate what looked simple from the literature, and emerged from Nature's survey of 1,576 researchers wasn't able to. Then I had a crisis of confidence, and then I learned that

The challenge is not to eliminate problems with reproducibility in

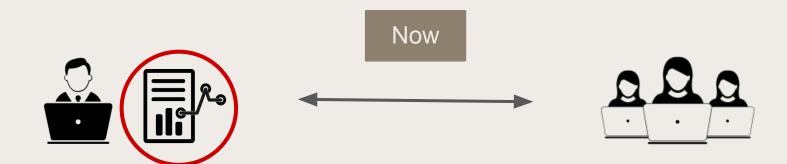
cancer biology³, found rates of around 40% and 10%, respectively. Our the wast majority of researchers in our survey had failed to reproduce survey respondents were more optimistic: 73% said that they think that an experiment, less than 20% of respondents said that they had ever at least half of the papers in their field can be trusted, with physicists and been contacted by another researcher unable to reproduce their work (see 'A 'crisis' in numbers'). Our results are strikingly similar to another The results capture a confusing snapshot of attitudes around these online survey of nearly 900 members of the American Society for issues, says Arturo Casadevall, a microbiologist at the Johns Hopkins Cell Biology (see go.nature.com/kbzs2b). That may be because such Bloomberg School of Public Health in Baltimore, Maryland. "At the conversations are difficult. If experimenters reach out to the original current time there is no consensus on what reproducibility is or should researchers for help, they risk appearing incompetent or accusatory, or

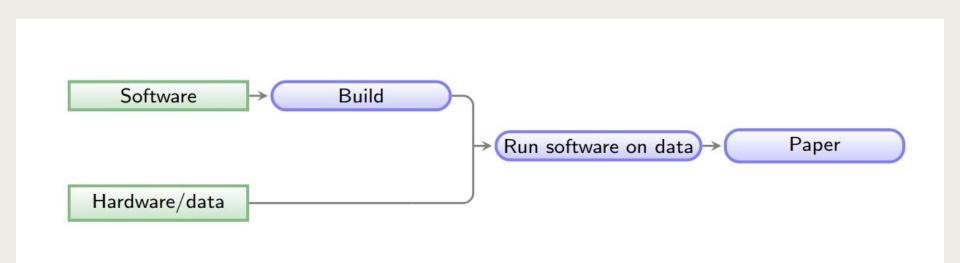
A minority of respondents reported ever having tried to publish

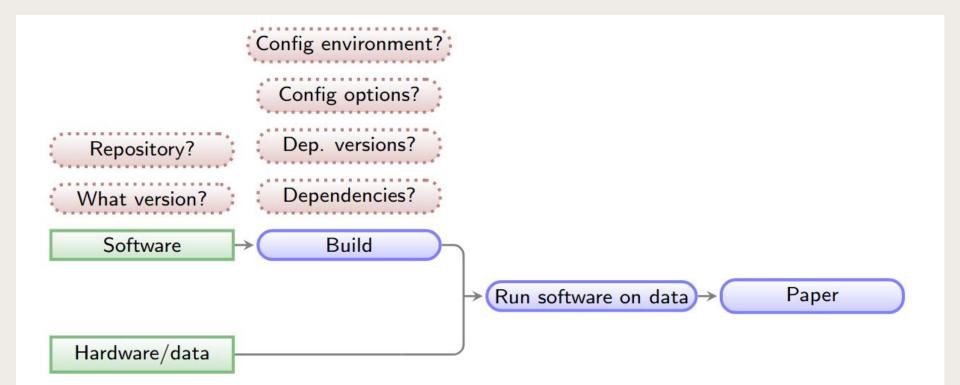


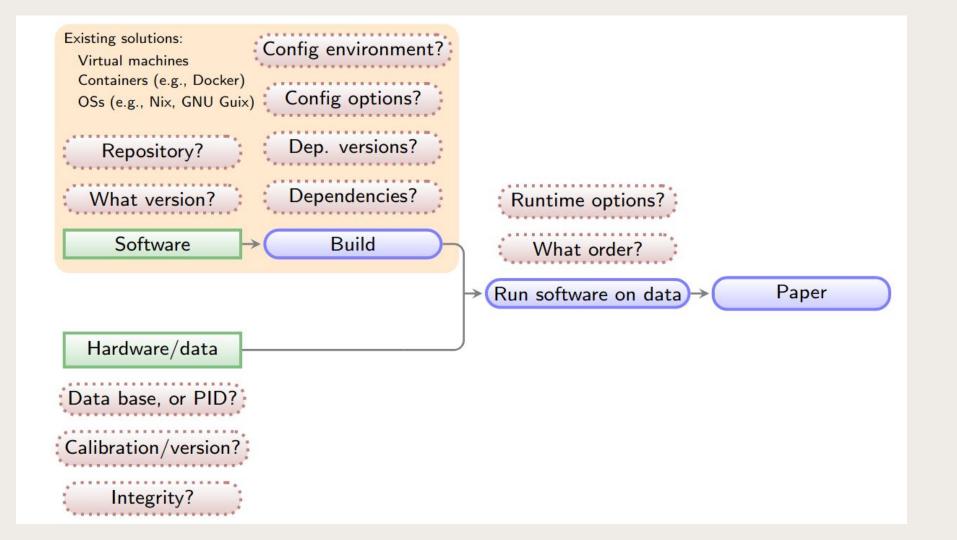
452 | NATURE | VOL 533 | 26 MAY 2016











Software dependencies

The objective is to make your analysis reproducible (by you+everyone)

An analysis may require different packages

- Load and process data
- Visualization
- Statistical analysis
- Advance processing
- ...







Software dependencies

The objective is to make your analysis reproducible (by you+everyone)

Software dependencies can change over time

An analysis that works today most probably will not work in a few years time. For example $py2 \rightarrow py3$.







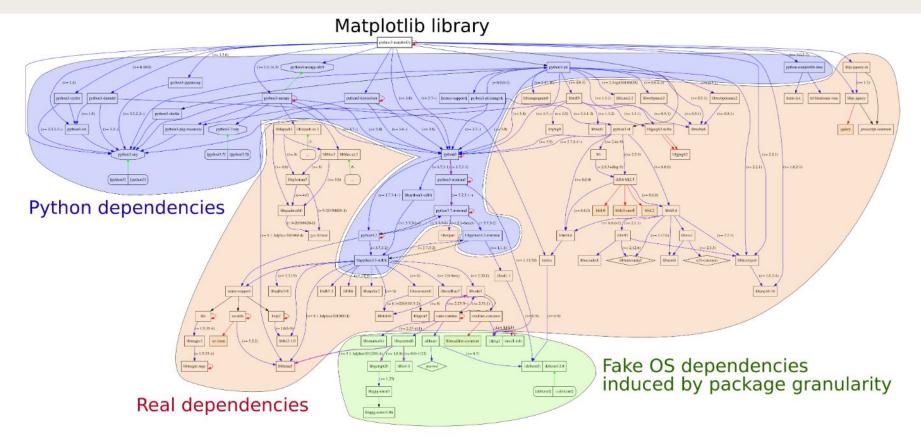


Fig. 1. Transitive dependencies of the software environment required by a simple "import matplotlib" command in the Python 3 interpreter.

Tracking all dependencies is very hard!

Some options:

- Write in the README the list of dependencies and versions
- Use a package manager (pip, Packrat, conda, ...)
 - Explicitly fix the dependencies and versions
 - Incorporates a way to install them
- Containers (singularity/docker)
 - Fixed "virtual machine", like a self-contained computer
 - Almost like a black box. Cannot be modified



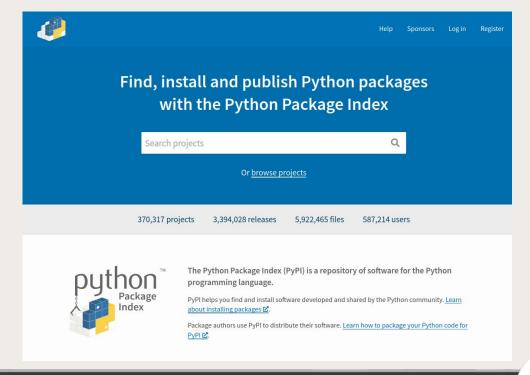


pip

pip is the <u>package</u> <u>installer</u> for Python.
You can use pip to install packages from the <u>Python Package</u> <u>Index</u> and other indexes.

pip

https://pypi.org/







Conda / Anaconda / miniconda

Why conda

What is conda?

Conda is a package manager used in scientific computing. It provides scientific libraries and dependencies.

Why conda?

- Manage the software for a project
- Can have different versions for each project
- You create virtual environments, encapsulated and reproducible





Conda

Conda

https://docs.conda.io/projects/conda/en/latest/index.html

Package, dependency and environment management for any language---Python, R, Ruby, Lua, Scala, Java, JavaScript, C/ C++, FORTRAN





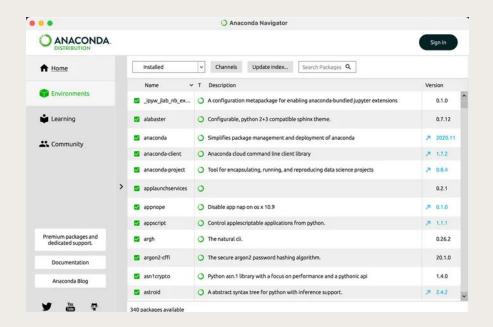


Anaconda

A private company providing software and package management solutions

Anaconda distribution

https://www.anaconda.com/products/distribution







conda-forge

A community-led collection of recipes, build infrastructure and distributions for the conda package manager

Package repository

https://conda-forge.org/

https://github.com/conda-forge





anaconda.org

Search available packages

https://anaconda.org/



Where packages, notebooks, projects and environments are shared.

SEARCH PACKAGES

Q Search Anaconda.org





Efficiency: miniconda mamba

miniconda

A lightweight version of conda. Very easy to install

https://docs.conda.io/en/latest/miniconda.html

mamba

A very fast dependency solver. Change conda → mamba

https://anaconda.org/conda-forge/mamba

conda install -c conda-forge mamba





Demo conda